Computers in Biology and Medicine 198 (2025) 111277

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Risk-adjusted training and evaluation for breast cancer detection

Dimitrios Bounias [a,b,*], Michael Baumgartner [a,c,h], Peter Neher [a,d,e], Balint Kovacs [a,b], Ralf Floca [a,f], Lorenz A. Kapsner [i], Jessica Eberle [i], Dominique Hadler [i], Frederik Laun [i], Sabine Ohlmeyer [i], Paul F. Jaeger [h,g], Michael Uder [i], Klaus H. Maier-Hein [a,b,c,d,e,j], Sebastian Bickelhaupt [i]

[a] German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[b] Medical Faculty Heidelberg, Heidelberg University, Im Neuenheimer Feld 672, 69120 Heidelberg, Germany
[c] Faculty of Mathematics and Computer Science, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
[d] German Cancer Consortium (DKTK), Partner Site Heidelberg, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[e] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Im Neuenheimer Feld 400, 69120 Heidelberg, Germany
[f] Heidelberg Institute of Radiation Oncology (HIRO), National Center for Radiation Research in Oncology (NCRO), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[g] German Cancer Research Center (DKFZ) Heidelberg, Interactive Machine Learning Group, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[h] German Cancer Research Center (DKFZ), Helmholtz Imaging, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany
[i] Institute of Radiology, Uniklinikum Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Ulmenweg 18, 91054 Erlangen, Germany
[j] National Center for Tumor Diseases (NCT), Heidelberg University Hospital (UKHD) and German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, 69120 Heidelberg, Germany

ARTICLE INFO

Keywords:
Object detection
Machine learning
Breast cancer
Medical imaging

ABSTRACT

Breast cancer detection, and broadly medical object detection, revolves around discovering and rating lesions. One of the most common ways of measuring performance is FROC (Free-response Receiver Operating Characteristic), which calculates sensitivity at predefined thresholds of false positives per case. However, depending on the clinical context, not all lesions might be of equivocal impact on the long-term outcome of a patient. Some lesions missed e.g. in screening might be detected in the subsequent screening round without impacting the clinical prognosis, whilst missing others might significantly detoriate prognosis and treatment pathways. It is therefore desirable to develop and include consideration of clinical prognosis/risk imbalance in the way machine learning models are developed and evaluated. In this work, we propose risk-adjusted FROC (raFROC), an adaptation of FROC that constitutes a first step on reflecting the underlying clinical need more accurately. Experiments on two independent breast magnetic resonance imaging (MRI) datasets with a total of 1535 lesions in 1735 subjects showcase the clinical potential of the proposed metric and its advantages over traditional evaluation methods. Additionally, by utilizing a risk-adjusted adaptation of focal loss (raFocal) we are able to improve the raFROC results and patient-level performance of nnDetection, at no expense of the regular FROC.

1. Introduction

In automatic diagnosis systems, discovery and rating of suspicious structures in patient images is commonly formulated as an object detection task [1–4]. Free-response Receiver Operating Characteristic (FROC), which is calculated as the mean of sensitivity at predefined numbers of false positives per case, is the standard measure of performance in medical object detection [1,2,5–7].

FROC and other detection metrics like mean Average Precision (mAP) [8], as well as the losses used to train machine learning (ML) models, treat all objects as equally important. However, depending on the clinical context and setting the diagnostic measure is embedded in, missing some lesions might be of very high clinical impact to the patient as it can significantly influence prognosis, whilst missing other lesions (postponing the detection in time but not in terms of prognosis), might not change either clinical outcome (e.g. survival) or clinical treatment

* Corresponding author at: German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany.
E-mail address: dimitrios.bounias@dkfz-heidelberg.de (D. Bounias).

https://doi.org/10.1016/j.compbiomed.2025.111277
Received 27 June 2024; Received in revised form 6 October 2025; Accepted 30 October 2025
Available online 5 November 2025

pathways (e.g. if a lesion is growing slightly in between two screening examinations but stays within the setting of identical prognosis and treatment pathways) [9,10]. Additionally, if lesions of subordinate immediate clinical relevance are more prevalent, the model might become biased towards them during training. Therefore, there is a disconnect between the underlying clinical need and the way models are optimized and evaluated, which can hinder clinical performance, make measuring scientific progress harder, and hamper clinical translation of ML [11].

To address this gap between technical methodology and clinical objective, we propose an adapted version of FROC, named risk-adjusted FROC (*raFROC*), that can incorporate arbitrary risk functions and account for risk differences between objects, aligning evaluation metrics with clinical needs. Additionally, we introduce risk-adjusted focal loss (*raFocal*) to support risk-aware training. In this way, we aim to focus the model evaluation and trained more on lesions of higher prognostic risk in case of being malignant, by considering them more important.

We showcase the risk-adjusted approaches by leveraging two distinct breast cancer datasets, each characterized by variations in lesion size distributions and MR imaging protocols, totaling 1535 lesions across 1735 subjects. While the methods presented can accept arbitrary risk functions, for this study we used tumor size — a key factor in breast cancer prognosis and TNM staging [12] — as the risk function, due to its well-established relationship with mortality risk [9,13].

2. Related work

2.1. Free-response operating characteristic (FROC)

The FROC plot shows sensitivity (often referred as true positive fraction, TPF) at certain manually defined thresholds of false positives per case (typically false positive detections per image, FPPI) [14,15] and constitutes the basis for raFROC. Each predictions has an associated confidence, represented with x for false positive (FP) and y for true positive (TP) ratings. Using multiple confidence thresholds represented by ξ, FROC can be calculated using:

$$\begin{aligned} TPF(\xi) &= \sum_{j=1}^{M} \frac{I(y_j > \xi)}{T} \\ FPPI(\xi) &= \sum_{i=1}^{N} \frac{I(x_i > \xi)}{S} \end{aligned} \tag{1}$$

where I the indicator function, M is the total number of TP predictions, N is the total number of FP predictions, and T is the total number of ground truth objects, and S the total number of cases.

2.2. Using focal loss as the classification loss functions for object detection

Focal loss [16], an enhanced version of cross-entropy loss, was designed specifically to address class imbalance in object detection tasks, where the vast number of easy negatives can overwhelm the training process. It modifies the standard cross-entropy loss by adding a focusing parameter, γ, which adjusts the rate at which easy examples are down-weighted. By doing so, focal loss effectively reduces the relative loss for well-classified examples (putting the focus on harder, misclassified examples), helping to prevent the overwhelming majority of easy negatives from dominating the gradient. For one-stage detectors, focal loss has become a popular choice, however this is an area of active development and many adaptations exist. However, when it assign weights to objects, it is typically based on class distributions or difficulty rather than any risk-aware approach.

The definition we utilize for focal loss is as follows:

$$focal(p_t) = -\alpha(1 - p_t)^{\gamma} log(p_t) \tag{2}$$

where α and γ are the typical focal loss hyperparameters.

2.3. Handling imbalance and risk in medical imaging

In object detection, the non-medical Common Objects in Context (COCO) challenge [17] presents average precision results for small, medium, and large objects, alongside the total overall score, which is a size-stratified analysis and moves in the direction of considering object properties during evaluation. Subject-centered FROC [14] is useful when diagnostic performance relates to the total number of targets per subject, however treats all detections within a patient as equal. Finally, cost-sensitive learning [18] prioritize errors with higher cost and shares similar concerns with this study, however remains understudied in medical imaging.

3. Methods

3.1. Risk-adjusted FROC (raFROC)

To account for risk in raFROC, we propose an adaptation to the FROC metric. TP predictions and ground truth samples are weighted by a weight $w \in [0, 1]$ pertaining to the associated risk, thus resulting in a risk-adjusted sensitivity. FP predictions are in turn weighted by $(1 - w)$, due to the desire to minimize unneeded medical measures or biopsies in persons without a true positive condition, resulting in a risk-adjusted number of FPs (see Fig. 1). A TP prediction that has double the risk of another one will be given double the weight during evaluation, or vice versa for a FP. Since FROC is understood as sensitivity at manually defined thresholds of FPs per case, raFROC can be understood as approximating high-risk object sensitivity at manually defined thresholds of low-risk FPs per case. Similarly to FROC, the final raFROC score is calculated by averaging the weighted sensitivities at all pre-defined FPPI thresholds [6,19,20]. As such, the equations become:

$$\begin{aligned} Risk\text{-}adjusted\ TPF(\xi) &= \sum_{j=1}^{M} \frac{w_i \cdot I(y_j > \xi)}{T} \\ Risk\text{-}adjusted\ FPPI(\xi) &= \sum_{i=1}^{N} \frac{(1 - w_i) \cdot I(x_i > \xi)}{S} \end{aligned} \tag{3}$$

3.2. Risk-adjusted focal loss function (raFocal)

We propose an adaptation to focal loss that accounts for risk (raFocal). To that end, we asymmetrically adjust the loss for a *risk-adapted* class, a class for which the risk calculation applies. Detection often uses sigmoid binary classification and thus each class can be adapted independently. Loss for other potential foreground classes is left untouched and the methodology below only applies to the risk-adapted class.

Let $y \in \{\pm 1\}$ indicate sufficient IoU (Intersection-over-Union) with a ground truth object (with y=−1 indicating the background class). The risk-adjusted weight $w_t \in [1, 2]$ of a prediction is then calculated by:

$$w_t = \begin{cases} 1 + w_{gt}, & \text{if } y = 1 \\ 2 - w_{pred}, & \text{if } y = -1 \end{cases} \tag{4}$$

where $w_{gt} \in [0, 1]$ is the risk of the matching (i.e. sufficient IoU overlap) ground truth box and $w_{pred} \in [0, 1]$ is the risk of the prediction box.

Thus, if the prediction is assigned to a ground truth box, the loss gets weighted by a value that increases as the risk of the ground truth box becomes larger. Conversely, for a prediction assigned to the background class, the value increases as the risk calculated for the prediction decreases.

The calculation for raFocal loss for the risk-adapted class finally becomes:

$$raFocal(p_t) = -w_t \alpha(1 - p_t)^{\gamma} log(p_t) \tag{5}$$

where $p_t \in [0, 1]$ is the model's estimated probability for the ground-truth class, α and γ are the focal loss hyperparameters.

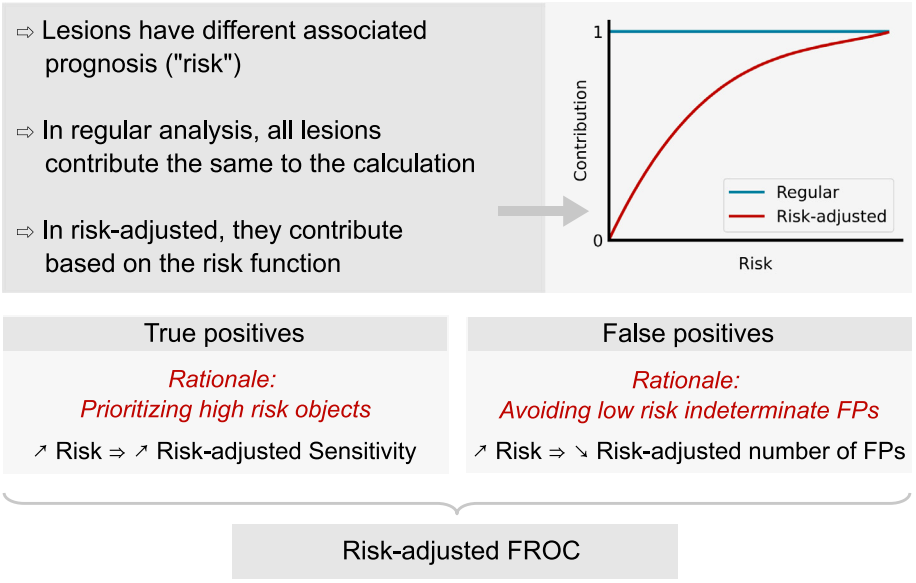

Fig. 1. Overview of the effects that different risk has on the calculation of the raFROC (risk-adjusted Free-response Receiver Operating Characteristic) metric.

3.3. Risk estimation

Based on an external study [9] involving 819,647 patients, we calculate a risk function for reported 15-year breast cancer mortality based on lesion size in millimeters. The choice of the external study was based on the fact that it was the largest study available, in an effort to minimize potential biases and thus expected to provide substantial robustness across breast cancer datasets. While this function pertains to breast cancer, similar functions can be constructed for other medical domains and disease entities with varying risks, potentially including other risk factors.

To calculate this function, we fit numbers available in the external study to a 3rd degree polynomial using numpy [21] (numpy.poly1d (numpy.polyfit(x, y, 3)), where x is an array of lesion sizes and y is an array of the respected values for mortality risk).

$$risk(size) = 2.28 \cdot 10^{-7} \cdot size^3 - 8.75 \cdot 10^{-5} \cdot size^2 + 1.23 \cdot 10^{-2} \cdot size + 1.37 \cdot 10^{-3} \quad (6)$$

Fig. 2 shows the risk function alongside the data distribution. Risk is further normalized to $[0, 1]$ by dividing by 0.641, the highest value in the $[0, 150]$ mm range of the study.

3.4. Data

3.4.1. In-house dataset

The ethics committee of Friedrich-Alexander-University (FAU) Erlangen-Nürnberg approved this retrospective study and waived the need for written informed consent. The dataset comprises 818 studies from 818 patients, including examinations acquired between November 2017 and January 2020, with an average age of 50.98 years. Among the dataset, 84 cases were classified as BI-RADS 4, 68 as BI-RADS 5 and 168 as BI-RADS 6 (BI-RADS: Breast Imaging-Reporting and Data System). Inclusion criteria were women undergoing clinically indicated breast MRI with a full multiparametric diagnostic protocol that included diffusion-weighted MRI (DW-MRI) sequences acquired with the b-values b=50, 750, 1500 s/mm^2, using 1.5T (Aera, Siemens Healthineers, Erlangen, Germany) and 3.0T (Skyra/Vida, Siemens Healthineers, Erlangen, Germany) clinical routine MRI scanners. The acquisitions had an average pixel spacing of 1.38 mm×1.38 mm and an average slice thickness of 4.4 mm.

Clinical indications for undergoing breast MRI included e.g. preoperative exclusion of multifocal disease, screening in women at high risk for breast cancer, exclusion of recurrent breast cancer, clarification of unclear findings in mammography, ultrasound or due to clinical complaints. The lesions were segmented as follows: All imaging data was transferred to a dedicated research workstation equipped with using open-source 3D Slicer Software [version 4.11] [22]. The built-in region draw function of the software was then used by a medical student with 2 years of experience in breast MRI research under the supervision of a board-certified radiologist with >10 years of experience to manually segment the lesions in the DW-MRI sequence using the high b-value (b=1500 s/mm^2) as target sequence. In case the target finding was not identifiable on the high b-value DW-MRI acquisition the intermediate b-value of 750 s/mm^2 was used for segmentation. Referencing of other sequences of the acquisition protocol (T2-weighted sequence, gadolinium-based contrast agent (GBCA) enhanced T1-weighted sequence) was allowed during the segmentation process.

The segmentations were performed along the inner border of the target finding for the full three-dimensional extension of the lesion. In case of multiple lesions in a single examination, all individual lesions clearly attributable (e.g., satellite lesions associated with malignant tumors/known multifocal disease) were segmented. Suspicious lesions were mostly histopathologically verified with imaging-guided biopsy. Biopsies were commonly not available in cases where the lesions were radiologically benign, known to be benign through follow-up examinations, or had been clarified through ultrasound examinations (e.g., cysts or fibroadenomas). Such lesions were classified as benign in this study. Histopathologically proven malignant lesions accounted for 618 lesions in 268 cases, while 1003 lesions in 373 cases were classified as benign to aid the model.

3.4.2. Duke-Breast-Cancer-MRI dataset

Publicly available [23–25], pre-operative dynamic contrast enhanced MRI of patients with biopsy-confirmed invasive breast cancer, including 917 cases with 917 malignant lesions (randomly selected independent test set: 230 images/230 lesions, Average age (available in 540 cases) was 52.88 years. Five cases excluded due to unclear preprocessing. The dataset includes axial breast MRI, in the prone position and the magnetic field strength of the scanners used was 1.5T/3.0T (using Skyra/Avanto/Trio/TrioTim, Siemens Healthineers, Erlangen, Germany and SIGNA HDx/SIGNA HDxt/SIGNA Excite/Optima MR450w, GE Medical Systems, Chicago, Illinois, USA) with an average pixel spacing

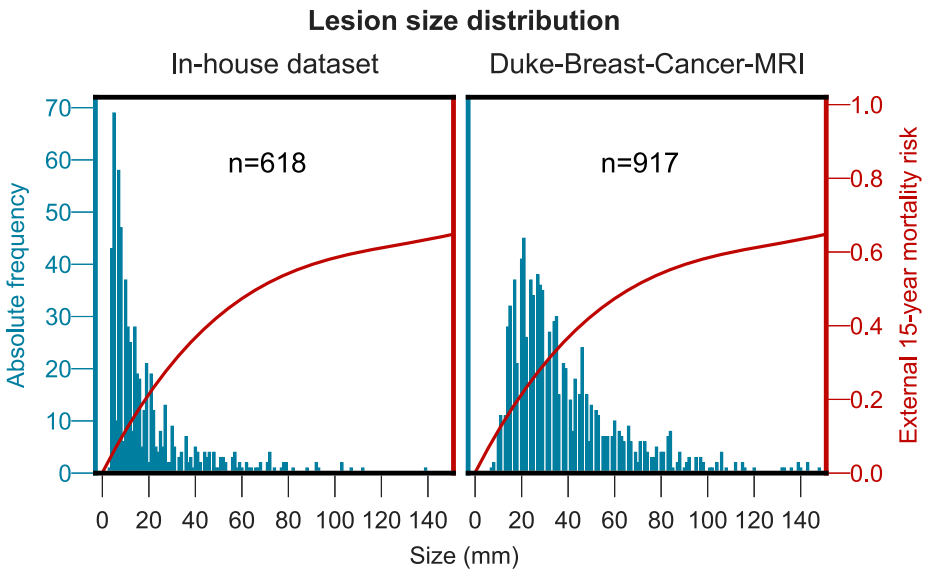

Fig. 2. Absolute frequencies of malignant tumor sizes, alongside the associated 15-year mortality risk of each size. The in-house dataset used for the evaluation is composed of diffusion-weighted MRI acquisitions with b-values 50,750,1500 s/mm^2, whilst the external Duke-Breast-Cancer-MRI dataset [23–25] is utilized as dynamic contrast enhanced (DCE) subtractions acquired before and after the injection of gadolinium-based contrast agents (GBCA), for further details see reference. Size is defined as the max axial detection box side.

of 0.72 mm × 0.72 mm and an average slice thickness of 1.15 mm. Three-dimensional bounding box position of each patient's primary lesion is included. For each patient, a non-fat saturated T1-weighted sequence, a fat-saturated gradient echo T1-weighted pre-contrast sequence, and at least three post-contrast sequences, after intravenous injection of a GBCA, are available. In the context of this study, we utilize subtraction images, in which the pre-contrast image was subtracted from the second post-contrast image.

3.5. Experimental design

3.5.1. Model

We use nnDetection [7] as the development framework, which is a self-configuring medical object detection framework built on Retina U-Net [26]. The self-configuring nature of the framework was utilized wherever possible in order to not inadvertently bias the model. Model parameters were optimized solely using focal loss and FROC on the in-house dataset. 8 epochs were found to perform best with 2500 batches per epoch and a learning rate of 0.01. The standard nnDetection augmentation pipeline was used. Focal loss hyperparameters α=0.75 and γ=1 were picked using the training set. For the asymmetric focal loss [27], γ was set to γ_+ (and γ_-=0). The ratio needed for focal-Tversky loss [28] was set to α=0.75/β=0.25 (as typically $\alpha + \beta = 1$ and $\alpha > \beta$ incentivizes false positive reduction). Hard negative mining with binary cross-entropy loss was also tested [7]. In order to assess the validity of the proposed loss and metric, we perform the following experiments.

3.5.2. Experiment 1: Empirically confirm shortcomings of FROC to motivate raFROC

We assume two hypothetical models, where one is equivalent to detecting only small breast lesions and another to detecting only large breast lesions. To achieve that, we pick a size threshold that splits the breast lesions into two buckets as evenly as possible. The two hypothetical models are able to predict breast lesions only from their respective bucket. We investigate whether there is a difference in patient-level sensitivity and risk-adjusted object-level sensitivity. This experiment is performed in the in-house dataset, which contains both malignant and benign cases.

3.6. Experiment 2: Qualitative comparison of proposed loss against focal loss and analysis of performance in different thresholds

Predictions from the focal loss trained model are compared visually to the ones from the model trained with raFocal loss. Additionally, different patterns in high confidence predictions from the two models are identified by utilizing different high confidence thresholds and comparing the number of false positives and true positives. Lastly, the numbers of true positive and false negative predictions of the two models, regardless of confidence (threshold 1%), are identified.

3.6.1. Experiment 3: Performance of proposed loss against focal loss using traditional evaluation metrics

We evaluate by means of aggregating 5-fold cross validation results for the in-house dataset and dedicated test set for the Duke dataset. The methods used are (1) FROC analysis for all datasets, (2) size-stratified FROC analysis for all datasets, and (3) Patient-level AUC (Area Under Curve) and patient-level AP (Average Precision) for the in-house dataset that has both pathological and non-pathological cases. Size-stratified analysis took place using the above-mentioned COCO method [17], where for a certain object size range, ground truth objects with size outside the range as well as non matching predictions with size outside the range are discarded prior to the calculation. 20 mm was chosen as the size range cut-off, because it constitutes the cut-off between T1 and T2 in the TNM staging system [12]. The results table also includes a comparison to asymmetric focal loss, focal-Tversky loss, and hard negative mining with binary cross-entropy loss.

3.6.2. Experiment 4: Performance of proposed loss against focal loss using the proposed evaluation metric

The hypothesis is that raFocal will perform better than focal loss, especially for the in-house dataset where there is an abundance of small lesions. We are also looking to gain insights on the risk-adjusted performance using the evaluation performed with the traditional methods and showcase that raFROC is a simpler approach with less overhead and pitfalls. The results table also includes a comparison to asymmetric focal loss, focal-Tversky loss, and hard negative mining with binary cross-entropy loss.

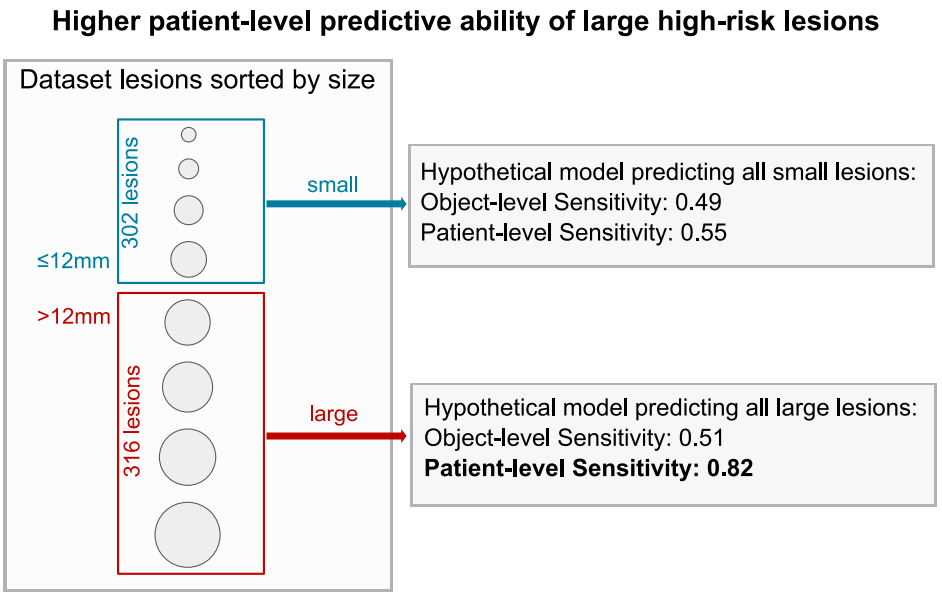

Fig. 3. Summary of the different predictive ability of lesion sizes using the in-house dataset. All lesions are divided into two buckets depending on their size and two hypothetical models are assumed that predict only from their respective bucket. Multiple lesions can belong to a single patient. Large lesions achieve higher sensitivity in predicting malignancy.

4. Results

4.1. Experiment 1: Empirically confirm shortcomings of FROC to motivate raFROC

A size threshold of 12 mm was chosen, in order to split the ground truth malignant lesions of the in-house dataset into two buckets as evenly as possible, which resulted in 302 malignant breast lesions in the small lesion size bucket and 316 in the large one. The resulting size range in the small bucket was 2.73–11.88 mm (7.13 mm average/6.84 mm median) and in the large 12.30–139.45 mm (30.35 mm average/21.88 mm median). We assume two hypothetical models, where the first one detects only all of the small lesions and the second only all the large ones. These models have similar object-level sensitivities, 0.49 and 0.51, as they are able to detect a similar amount of malignant breast lesions. However, the diagnostic value of the two models differs, as smaller lesions are often accompanied by larger ones in the same patient and because there can be multiple small lesions per patient. The small lesion model can achieve a patient-level sensitivity of 0.55, compared to 0.82 for the large lesion one. The risk-adjusted object-level sensitivity that raFROC uses would be 0.24 and 0.76, respectively. Fig. 3 summarizes the results.

4.2. Experiment 2: Qualitative comparison of proposed loss against focal loss and analysis of performance in different thresholds

Fig. 4 contains case examples between models trained with focal loss and raFocal loss. In particular, lesions with higher risk — here larger size — were more likely to be caught by the raFocal model. Table 1 shows that the model trained with raFocal loss tend to find more high risk lesions with a high confidence score and having less low risk false-positives, at the expense of also being more confident with high risk-appearing false positives.

With a threshold of 1% we calculate an estimate of the number of overall discovered ground truth lesions by the two models. In the in-house dataset, the number of found lesions are 493/618 for the model trained with focal loss and 514/618 for the raFocal model. In Duke, there are 212/230 lesions found for the model with focal loss and 216/230 for the raFocal.

4.3. Experiment 3: Performance of proposed loss against focal loss using traditional evaluation metrics

Fig. 5 shows the results of the FROC analysis. The performance metrics are aggregated in Table 2. For the in-house dataset, the raFocal model achieves 0.86 AUC and 0.77 AP compared to 0.84 AUC and 0.70 AP for the focal model ($p = 0.006$ significant difference between AUC scores using the DeLong method [29]) on the patient-level. The raFocal model performs moderately better at the object-level in lower thresholds. The higher sensitivity of the focal model at 1 FP/case in the 20+mm plot is misleading, as this happens for a very low prediction probability threshold (0.025) which is not relevant, as it is lower than the largest threshold used in the overall FROC plot at 8 FPs/case (0.039).

For the Duke dataset, the analysis shows improvement for raFocal in all FROC plots. However, upon closer inspection, most of the benefit in the overall FROC performance comes from the 20+mm range and the two plots follow a similar pattern. That is because the probability thresholds used for the overall FROC curve produce very similar number of 0–20 mm TPs. In fact, the largest difference is in the lowest threshold, where focal loss actually has more 0-20 mm TPs, 24 vs 20, something not visible in the 0-20 mm plot.

4.4. Experiment 4: Performance of proposed loss against focal loss using the proposed evaluation metric

The results (Fig. 6) indicate that raFocal shows improvements in raFROC compared to focal loss, especially for the in-house dataset, where there is more variance in the lesion sizes. Specifically, there is an increase from 0.60 raFROC to 0.65 for the in-house dataset and 0.88 to 0.91 for the Duke dataset. This is in accordance with the increase in the patient-level metrics observed and the slight increase in 0–20 mm lesion performance for the lower thresholds. The improvement raFocal shows in the Duke dataset stems from better large lesion (20 mm+) performance, as indicated by the COCO-style size-stratified analysis.

5. Discussion

In this study, we introduced risk-adjusted FROC (raFROC) and risk-adjusted focal loss (raFocal) to incorporate clinical context concerns

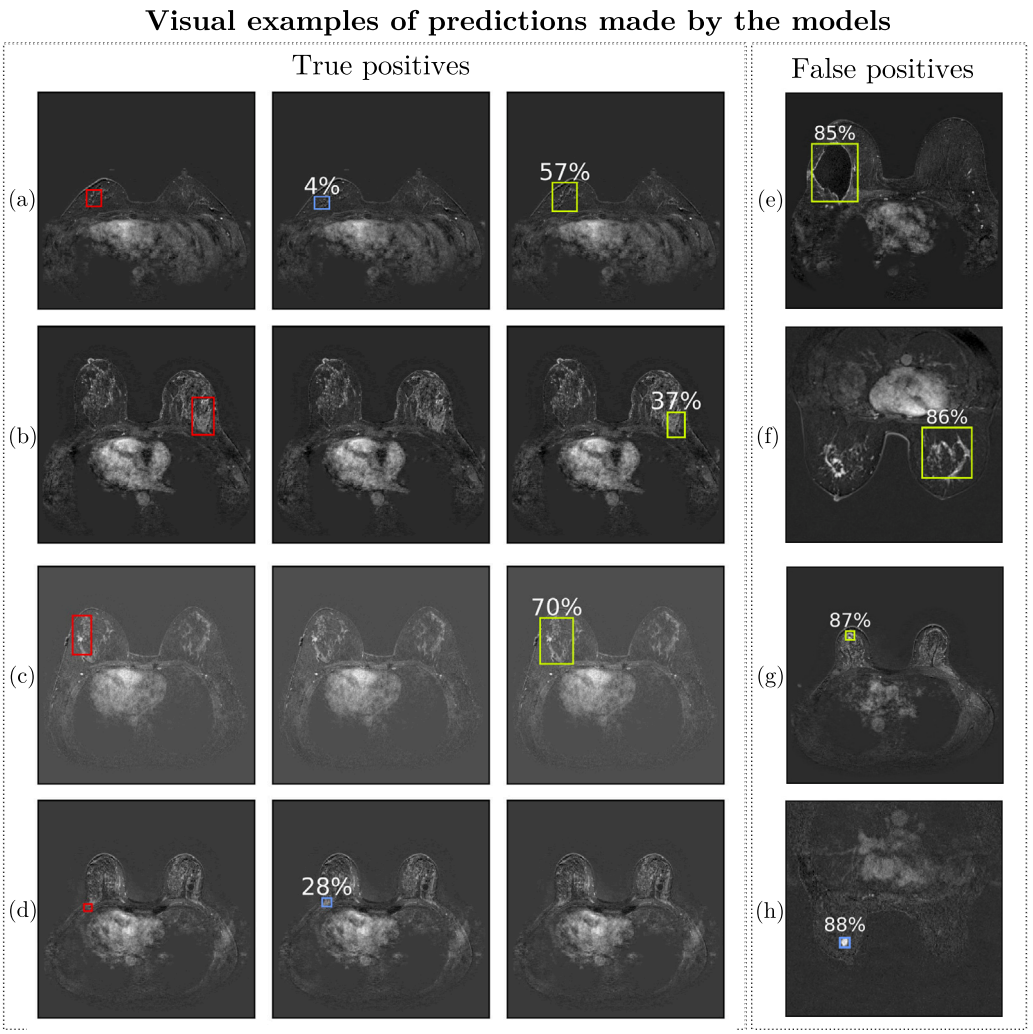

Fig. 4. Qualitative assessment using representative detection examples made by the model. Examples (a)–(d) are cases where the models differed in assessing a ground truth lesion; from left to right: ground truth, prediction (if any) from focal loss model, prediction (if any) from raFocal loss model. Examples (e)–(h) are cases where the models differed in confidently classifying a lesion as ground truth: yellow predictions are from the raFocal loss model, blue from the focal loss model. Images are from the external Duke-Breast-Cancer-MRI dataset (CC BY-NC 4.0 license) [23–25], which is utilized as dynamic contrast enhanced (DCE) subtractions acquired before and after the injection of gadolinium-based contrast agents (GBCA).

Table 1
Comparison of false positives (FP) and true positives (TP) for large and small objects across different detection thresholds (75%, 80%, 85%) using focal loss and risk-adjusted focal loss (raFocal). Results are reported for both the in-house dataset and the Duke dataset.

Dataset	Category	75%		80%		85%	
		Focal	raFocal	Focal	raFocal	Focal	raFocal
In-house	Large FP	**81**	98	**61**	77	**42**	55
	Large TP	109	109	104	**106**	89	**96**
	Small FP	427	**262**	349	**196**	241	**130**
	Small TP	**156**	126	**142**	119	**118**	102
Duke	Large FP	**22**	30	**16**	24	**7**	10
	Large TP	137	**146**	127	**136**	94	**116**
	Small FP	8	8	4	4	4	**3**
	Small TP	24	24	20	20	12	**16**

into object detection in medical imaging. Our methods were evaluated on two independent breast MRI datasets: an in-house DW-MRI dataset and a publicly available DCE-MRI dataset. The datasets were modified in order to represent two distinct different potential screening applications using abbreviated breast MRI. The in-house dataset was used to represent a non-contrast enhanced imaging strategy based on high b-value DW-MRI and the public dataset on representing a GBCA-enhanced breast MRI approach.

Our experimental results demonstrated an increase in performance using raFocal across both datasets. This improvement was evident in patient-level and object-level performance, particularly in risk-adjusted evaluations and the patient-level AUC. Incorporating risk adjustment (in the sense of long-term survival prognosis, in case a lesion turns out to be malignant) into training led to more accurate identification of lesions with higher prognostic burden for the patients. We also observed that larger lesions have a vastly better patient-level sensitivity.

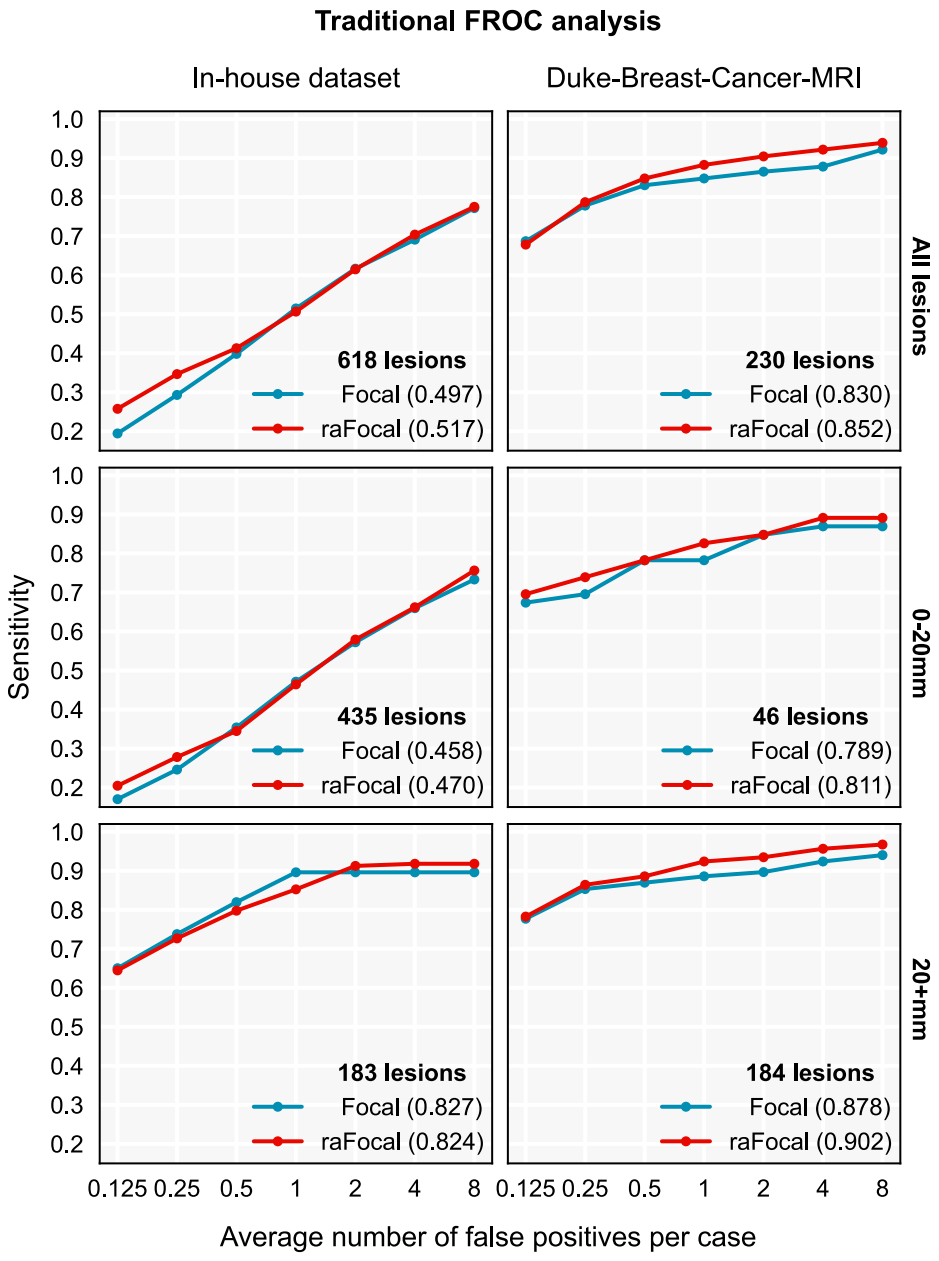

Fig. 5. Performance of the proposed loss function (raFocal - risk-adjusted focal loss) against focal loss using traditional evaluation metrics. The first row is Free-response Receiver Operating Characteristic (FROC), while the next two rows comprise the size-stratified FROC analysis with the method used in COCO [17].

Table 2
Performance of the proposed loss (raFocal, risk-adjusted focal loss) against standard focal loss, asymmetric focal loss, focal-Tversky loss, and hard negative mining with binary cross-entropy. The comparison takes place across an diffusion-weighted MRI in-house dataset and composed dynamic contrast enhanced subtractions from Duke-Breast-Cancer-MRI [23–25]. Metrics include patient-level area under the receiver operating characteristic curve (AUC), patient-level average precision (AP), object-level free-response operating characteristic (FROC), and object-level risk-adjusted FROC (raFROC). Significance testing was performed for AUC, using the DeLong method [29], which indicated that the difference was significant ($p = 0.006$).

Model loss	In-house dataset				Duke	
	AUC	AP	FROC	raFROC	FROC	raFROC
raFocal	**0.86**	**0.77**	**0.517**	**0.645**	**0.852**	**0.906**
Focal	0.84	0.70	0.497	0.598	0.830	0.878
Asymmetric Focal	0.85	0.72	0.501	0.599	0.843	0.896
Focal-Tversky	0.65	0.45	0.160	0.263	0.412	0.431
Hard Negative Mining	0.80	0.64	0.457	0.569	0.834	0.884

The raFROC analysis provided a nuanced evaluation that considered the size-association risk, further supporting these findings.

The raFROC approach offers a more clinically relevant analysis compared to traditional FROC and COCO-style stratified analyses. While

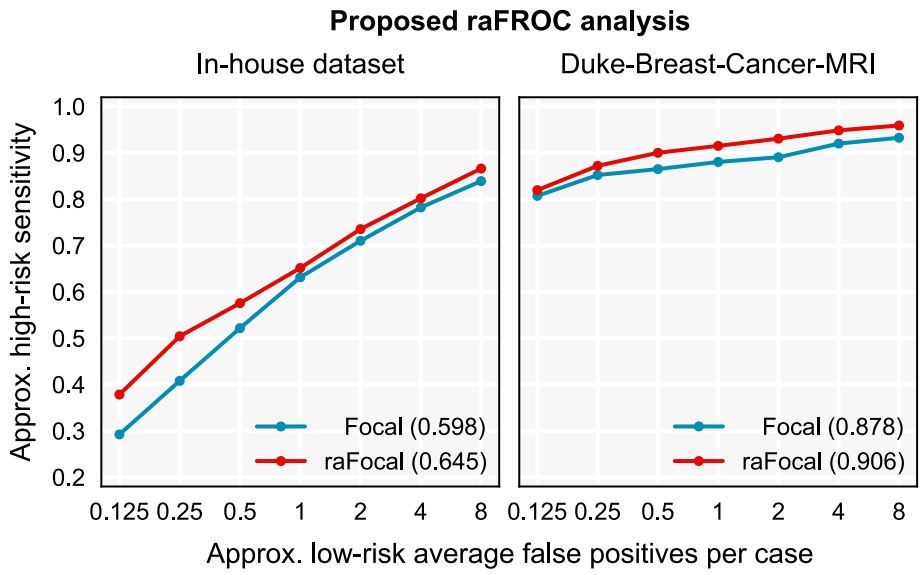

Fig. 6. Performance of the proposed loss function (raFocal - risk-adjusted focal loss) against focal loss using the proposed evaluation metric, raFROC (risk-adjusted Free-response Receiver Operating Characteristic), in a diffusion-weighted MRI in-house dataset and the public Duke-Breast-Cancer-MRI dataset [23–25], which is utilized as dynamic contrast enhanced (DCE) subtractions acquired before and after the injection of gadolinium-based contrast agents (GBCA).

both the presented implementation of raFROC and the COCO-style size-stratified analysis are based on size calculations, the size-stratified analysis can be misleading due to the need for arbitrary size thresholds and the assumption of uniform risk within each size range. Additionally, COCO-style analysis involves complexities such as managing multiple prediction probability thresholds for each size range, which can obscure true model performance. In contrast, raFROC incorporates clinical prognosis directly into the evaluation metric, providing a unified and straightforward measure of model performance.

For clinical practice, this means that the model can potentially aid in the detection and accurate characterization of breast cancer lesions, reducing concerns about missing important findings and overdiagnosis. For radiologists, automated detection systems in the future might support highlighting lesions that require immediate attention, potentially reducing time spent on findings with limited clinical relevance and focusing efforts on critical cases.

Adequate loss functions are essential for representing the addressed problem and data, accounting for any imbalances. Standard loss functions like focal loss emphasize harder examples but do not consider the clinical impact of different lesion types. The proposed raFocal loss adjusts weights based on the potential prognostic downstream risk in case a lesion would have been malignant, enhancing the detection of clinically significant lesions. Our results indicate that raFocal outperformed focal loss in both datasets, particularly in improving the detection of larger lesions with a potential prognostically worse risk in case they turn out to be malignant in the Duke dataset, and significantly enhancing patient-level AUC and AP in the in-house dataset. The raFocal loss indirectly incorporates both patient-level performance and object-level performance, which was previously missing during model training.

This work bears similarities to net benefit [10], where the goal is to detect malignancies while minimizing unnecessary interventions. Net benefit establishes an exchange rate between identifying true positives and reducing false positives, ensuring a balanced approach in developing screening programs. By integrating clinical relevance, such as treatment success and survival prognosis, we utilized net benefit concepts to help optimize medical object detection for better patient outcomes.

There are several limitations to our study. First, we only used lesion size as a risk adjustment factor, while breast cancer prognosis is typically determined by a diverse spectrum of factors. This limitation means that the current risk adjustment is certainly not able to fully capture the complexity of breast cancer prognosis. Future research should aim to integrate a broader range of risk factors, such as individual risk factors (e.g., BRCA mutation carrier) and tumor genetic characteristics to enhance accuracy and applicability. As these risk factors were not available for all datasets, we focused on a basic, established factor associated with long-term patient outcomes. However, we showcased the potential of incorporating additional measures of risk and even the straightforward risk function utilized showed promising results. Further the risk described in our study characterizes the clinical downstream prognostic risk in case a lesion turns out to malignant. It is not a risk factor that characterizes the risk of a lesion itself to be malignant as such. This is important to consider as size itself cannot be sufficiently used to differentiate breast lesions into malignant and benign.

Secondly, our evaluation was confined to breast cancer datasets, and the generalizability to other cancers or medical conditions remains to be validated. Similarly, the underlying patient cohort did not reflect the characteristics of a screening population in terms of frequency of malignant findings and lesion characteristics. Further, the moderate performance of the neural network in the in-house dataset could be attributed to the high number of small benign lesions often present alongside malignant ones in non-contrast-enhanced MRI, making the dataset particularly challenging. In contrast, the Duke dataset contained larger malignant lesions, favoring the neural network's detection capabilities. Our work does not incorporate all types of current AI applications used in breast imaging [30–34] and we purposely utilized a self-configuring method without fine-tuning the model in order to not bias results.

Future work could focus on expanding the risk models to encompass a broader range of clinical factors and on validating these methods across diverse medical imaging datasets or different cancer types, like lung cancer, as well as incorporation of more patient-level concerns like treatment response. Screening datasets, while potentially more difficult to acquire, could reveal an even higher relevance of risk-adjusted methods. Federated learning [35] and the large datasets it involves can potentially benefit from such an approach, as population differences can present differences in risk characteristics. Further exploration into the integration of risk-adjusted methods into real-time clinical decision support systems could assess their impact on clinical workflows and patient outcomes. By addressing these points, the effectiveness of medical object detection models can be enhanced, ensuring they are better aligned with clinical needs and ultimately improving patient care.

6. Conclusions

This work showcases a first step on how to realize risk-adjusted model training and validation in medical object detection. Accounting for clinical risk and outcome is very important in a medical diagnosis setting compared to other domains, because it allows balancing the trade-off between false positive findings and missing pathologies, significantly influencing the clinical outcome of the individual patient. Further, this work paves the way on helping models identify lesions of high clinical relevance, with impact on the clinical prognosis. The method presented brings model evaluation and training loss closer to that need and is a better approach than size stratification. Source code for raFROC and raFocal loss is publicly available at https://github.com/MIC-DKFZ/risk-adjusted-training-and-evaluation.

CRediT authorship contribution statement

Dimitrios Bounias: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michael Baumgartner:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis. **Peter Neher:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis. **Balint Kovacs:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis. **Ralf Floca:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis. **Lorenz A. Kapsner:** Writing – review & editing, Formal analysis, Data curation. **Jessica Eberle:** Writing – review & editing, Validation, Data curation. **Dominique Hadler:** Writing – review & editing, Validation, Data curation. **Frederik Laun:** Writing – review & editing, Validation, Data curation. **Sabine Ohlmeyer:** Writing – review & editing, Validation, Data curation. **Paul F. Jaeger:** Writing – review & editing, Methodology. **Michael Uder:** Writing – review & editing, Resources, Project administration. **Klaus H. Maier-Hein:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis. **Sebastian Bickelhaupt:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partially supported by the “ForTra gGmbH für Forschungsstransfer der Else Kröner-Fresenius-Stiftung”.

References

[1] F. Ayatollahi, S.B. Shokouhi, R.M. Mann, J. Teuwen, Automatic breast lesion detection in ultrafast DCE-MRI using deep learning, Med. Phys. 48 (10) (2021) 5897–5907, http://dx.doi.org/10.1002/mp.15156.

[2] G. Maicas, A.P. Bradley, J.C. Nascimento, I. Reid, G. Carneiro, Deep reinforcement learning for detecting breast lesions from DCE-MRI, in: L. Lu, X. Wang, G. Carneiro, L. Yang (Eds.), Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics, in: Advances in Computer Vision and Pattern Recognition, Springer International Publishing, Cham, 2019, pp. 163–178, http://dx.doi.org/10.1007/978-3-030-13969-8_8.

[3] G. Maicas, G. Carneiro, A.P. Bradley, J.C. Nascimento, I. Reid, Deep reinforcement learning for active breast lesion detection from DCE-MRI, in: M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D.L. Collins, S. Duchesne (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2017, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2017, pp. 665–673, http://dx.doi.org/10.1007/978-3-319-66179-7_76.

[4] A. Gubern-Mérida, R. Martí, J. Melendez, J.L. Hauth, R.M. Mann, N. Karssemeijer, B. Platel, Automated localization of breast cancer in DCE-MRI, Med. Image Anal. 20 (1) (2015) 265–274, http://dx.doi.org/10.1016/j.media.2014.12.001.

[5] A. Reinke, M.D. Tizabi, C.H. Sudre, M. Eisenmann, T. Rädsch, M. Baumgartner, L. Acion, M. Antonelli, T. Arbel, S. Bakas, P. Bankhead, A. Benis, M.J. Cardoso, V. Cheplygina, E. Christodoulou, B. Cimini, G.S. Collins, K. Farahani, B. van Ginneken, B. Glocker, P. Godau, F. Hamprecht, D.A. Hashimoto, D. Heckmann-Nötzel, M.M. Hoffman, M. Huisman, F. Isensee, P. Jannin, C.E. Kahn, A. Karargyris, A. Karthikesalingam, B. Kainz, E. Kavur, H. Kenngott, J. Kleesiek, T. Kooi, M. Kozubek, A. Kreshuk, T. Kurc, B.A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A.L. Martel, P. Mattson, E. Meijering, B. Menze, D. Moher, K.G.M. Moons, H. Müller, B. Nichyporuk, F. Nickel, M.A. Noyan, J. Petersen, G. Polat, N. Rajpoot, M. Reyes, N. Rieke, M. Riegler, H. Rivaz, J. Saez-Rodriguez, C.S. Gutierrez, J. Schroeter, A. Saha, S. Shetty, M. van Smeden, B. Stieltjes, R.M. Summers, A.A. Taha, S.A. Tsaftaris, B. Van Calster, G. Varoquaux, M. Wiesenfarth, Z.R. Yaniv, A. Kopp-Schneider, P. Jäger, L. Maier-Hein, Common limitations of image processing metrics: A picture story, 2022, URL arXiv:2104.05642[cs,eess].

[6] A.A.A. Setio, A. Traverso, T. de Bel, M.S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M.E. Fantacci, B. Geurts, R. van der Gugten, P.A. Heng, B. Jansen, M.M. de Kaste, V. Kotov, J.Y.-H. Lin, J.T. Manders, A.S. nora Mengana, J.C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C.M. Schaefer-Prokop, E.T. Scholten, L. Scholten, M.M. Snoeren, E.L. Torres, J. Vandemeulebroucke, N. Walasek, G.C. Zuidhof, B. van Ginneken, C. Jacobs, Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge, Med. Image Anal. 42 (2017) 1–13, http://dx.doi.org/10.1016/j.media.2017.06.015.

[7] M. Baumgartner, P.F. Jäger, F. Isensee, K.H. Maier-Hein, nnDetection: A self-configuring method for medical object detection, in: M. de Bruijne, P.C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, C. Essert (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, Springer International Publishing, Cham, 2021, pp. 530–539.

[8] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2010) 303–338.

[9] V. Sopik, S.A. Narod, The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer, Breast Cancer Res. Treat. 170 (3) (2018) 647–656, http://dx.doi.org/10.1007/s10549-018-4796-9.

[10] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ (2016) i6, http://dx.doi.org/10.1136/bmj.i6.

[11] L. Maier-Hein, A. Reinke, P. Godau, M.D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, et al., Metrics reloaded: recommendations for image analysis validation, Nature Methods (2024) 1–18.

[12] G. Cserni, E. Chmielik, B. Cserni, T. Tot, The new TNM-based staging of breast cancer, Virchows Arch. 472 (5) (2018) 697–703, http://dx.doi.org/10.1007/s00428-018-2301-9.

[13] C. Verschraegen, V. Vinh-Hung, G. Cserni, R. Gordon, M.E. Royce, G. Vlastos, P. Tai, Modeling the effect of tumor size in early breast cancer, Ann. Surg. 241 (2) (2005).

[14] A.I. Bandos, H.E. Rockette, D. Gur, Subject-centered free-response ROC (FROC) analysis, Med. Phys. 40 (5) (2013) 051706.

[15] D.C. Edwards, M.A. Kupinski, C.E. Metz, R.M. Nishikawa, Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model, Med. Phys. 29 (12) (2002) 2861–2870.

[16] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 740–755, http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[18] C. Elkan, The foundations of cost-sensitive learning, in: International Joint Conference on Artificial Intelligence, vol. 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[19] M.U. Dalmış, S. Vreemann, T. Kooi, R.M. Mann, N. Karssemeijer, A. Gubern-Mérida, Fully automated detection of breast cancer in screening MRI using convolutional neural networks, J. Med. Imaging 5 (01) (2018) 1, http://dx.doi.org/10.1117/1.JMI.5.1.014502.

[20] M. Niemeijer, M. Loog, M.D. Abràmoff, M.A. Viergever, M. Prokop, B. van Ginneken, On combining computer-aided detection systems, IEEE Trans. Med. Imaging 30 (2) (2011) 215–223, http://dx.doi.org/10.1109/TMI.2010.2072789.

[21] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J.F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T.E. Oliphant, Array programming with NumPy, Nature 585 (7825) (2020) 357–362, http://dx.doi.org/10.1038/s41586-020-2649-2.

[22] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J.V. Miller, S. Pieper, R. Kikinis, 3D slicer as an image computing platform for the quantitative imaging network, Magn. Reson. Imaging 30 (9) (2012) 1323–1341, http://dx.doi.org/10.1016/j.mri.2012.05.001.

[23] A. Saha, M.R. Harowicz, L.J. Grimm, J. Weng, E.H. Cain, C.E. Kim, S.V. Ghate, R. Walsh, M.A. Mazurowski, Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations, 2022, http://dx.doi.org/10.7937/TCIA.E3SV-RE93, Type: dataset.

[24] A. Saha, M.R. Harowicz, L.J. Grimm, C.E. Kim, S.V. Ghate, R. Walsh, M.A. Mazurowski, A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features, Br. J. Cancer 119 (4) (2018) 508–516, http://dx.doi.org/10.1038/s41416-018-0185-8.

[25] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The cancer imaging archive (TCIA): Maintaining and operating a public information repository, J. Digit. Imaging 26 (6) (2013) 1045–1057, http://dx.doi.org/10.1007/s10278-013-9622-7.

[26] P.F. Jaeger, S.A.A. Kohl, S. Bickelhaupt, F. Isensee, T.A. Kuder, H.-P. Schlemmer, K.H. Maier-Hein, Retina U-net: Embarrassingly simple exploitation of segmentation supervision for medical object detection, in: A.V. Dalca, M.B. McDermott, E. Alsentzer, S.G. Finlayson, M. Oberst, F. Falck, B. Beaulieu-Jones (Eds.), Proceedings of the Machine Learning for Health NeurIPS Workshop, in: Proceedings of Machine Learning Research, vol. 116, PMLR, 2020, pp. 171–183.

[27] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 82–91.

[28] N. Abraham, N.M. Khan, A novel focal tversky loss function with improved attention u-net for lesion segmentation, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, IEEE, 2019, pp. 683–687.

[29] X. Sun, W. Xu, Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves, IEEE Signal Process. Lett. 21 (11) (2014) 1389–1393, http://dx.doi.org/10.1109/LSP.2014.2337313.

[30] M. Gagliardi, T. Ruga, E. Zumpano, E. Vocaturo, Breast cancer classification via deep learning approaches, SPAST Rep. 1 (4) (2024).

[31] Y. Jiang, A.V. Edwards, G.M. Newstead, Artificial intelligence applied to breast MRI for improved diagnosis, Radiology 298 (1) (2021) 38–46.

[32] N. Eisemann, S. Bunk, T. Mukama, H. Baltus, S.A. Elsner, T. Gomille, G. Hecht, S. Heywang-Köbrunner, R. Rathmann, K. Siegmann-Luz, et al., Nationwide real-world implementation of AI for cancer detection in population-based mammography screening, Nature Med. (2025) 1–8.

[33] M. Gagliardi, T. Ruga, E. Vocaturo, E. Zumpano, Predictive analysis for early detection of breast cancer through artificial intelligence algorithms, in: International Conference on Innovations in Computational Intelligence and Computer Vision, Springer, 2024, pp. 53–70.

[34] E. Vocaturo, E. Zumpano, Artificial intelligence approaches on ultrasound for breast cancer diagnosis, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2021, pp. 3116–3121.

[35] L. Caroprese, T. Ruga, E. Vocaturo, E. Zumpano, Federated learning applications for breast cancer, in: 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2023, pp. 4029–4034.